Parsimonious Bayesian Factor Analysis for modelling latent structures in spectroscopy data

Working Group on Statistical Learning

Alessandro Casa

Joint work with: Tom O'Callaghan & Brendan Murphy

School of Mathematics and Statistics
University College Dublin

alessandro.casa@ucd.ie

8th February 2021

A bit of context - Cows & Diet

- Consumers awareness is changing the food industry
  $\longrightarrow$ food quality, authentication, security and animal welfare

- Cattle feeding regimen:
  - outdoor pasture vs indoor total mixed ration
  - pasture feeding is regarded as more respectful and resulting products as more natural and healtier
  - enhanced beneficial nutrients and improved organolectic characteristics

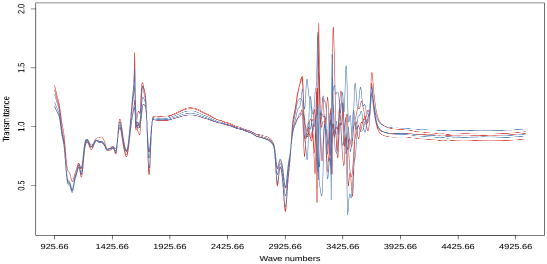
    $\downarrow$

  - Premium prices lead to food adulteration and fraud

- Methods to detect adulterant and to authenticate milk traceability are needed

A bit of context - Spectroscopy data

- Biomarkers-based methods have been proposed
  $\longrightarrow$ usually expensive and time consuming
- Vibrational spectroscopy techniques:
  - cheap, rapid and non-disruptive way to collect huge amount of data
  - used in food authenticity and in the dairy framework but way less explored to authenticate cow diets
  - How do they work?

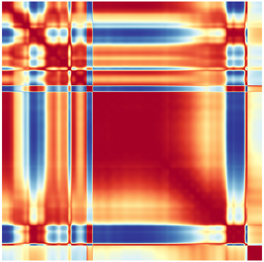


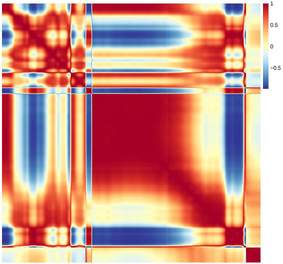
All that glitters is not gold

- Some challenges:
  - High-dimensionality
  - Peculiar correlation structures
- Aim:
  - Parsimonious representation
  - Proper wavelengths relationships reconstruction
  - Characterization of the phenomenon

    ↓

    Long story short: extract useful knowledge and insights





Latent variable models

- PLS and PCA: state of the art when analyzing spectral data
  $\longrightarrow$ regression and classification purposes

- Focus on Factor Analysis (FA):
  maps the data in lower-dimensional latent space while preserving the original information

- Two birds, one stone
  - dimensionality reduction...
  - with a focus on the covariance structure

Factor analysis in a nutshell

- Let $X = \{x_1, \dots, x_n\}$, with $x_i \in \mathbb{R}^p$, be the observed data. FA models each observation as follows

$$x_i = \mu + \Lambda u_i + \varepsilon_i, \qquad i = 1, \dots, n$$

  where $\varepsilon \sim \mathcal{N}_p(0, \Psi)$, $\Lambda \in \mathbb{R}^{p \times K}$ and $u_i \in \mathbb{R}^K$ the factor loading matrix and the factor scores with K the number of factors

- Standard distributional assumptions (wlog $\mu = 0$) lead to
  - $(x_i | u_i) \sim \mathcal{N}_p(\Lambda u_i, \Psi)$
  - $x_i \sim \mathcal{N}_p(0, \Sigma = \Lambda\Lambda^T + \Psi)$

- Focus on Λ as solely responsible for correlation reconstruction

Factor analysis in a nutshell

- In matrix form

$$\begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,p} \end{pmatrix} = \begin{pmatrix} \lambda_{1,1} & \cdots & \lambda_{1,K} \\ \vdots & \ddots & \vdots \\ \lambda_{p,1} & \dots & \lambda_{p,K} \end{pmatrix} \begin{pmatrix} u_{i,1} \\ \vdots \\ u_{i,K} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \vdots \\ \varepsilon_{i,p} \end{pmatrix}$$

- Lot of works focusing on sparse estimation of Λ to enhance interpretability and to detect uninformativeness

What about redundancy?

$\downarrow$

little to none attention given to situations where observed features are providing the same information

Factor analysis with redundant variables

- Idea: search for variable clustering structures
  $\longrightarrow$ modify standard FA by allowing some variables to be mapped by means of the same loadings

- Proposed model is defined as follows

$$\begin{aligned} x_i &= Z\Lambda_c u_i + \varepsilon_i \\ &= \tilde{\Lambda} u_i + \varepsilon_i, \qquad\qquad i = 1, \ldots, n \end{aligned}$$

  $Z \in \mathbb{R}^{p \times G}$ and $\Lambda_c \in \mathbb{R}^{G \times K}$ and G the number of variable clusters

  - Z latent *allocation matrix*. Standard binary partition i.e. $z_{jg} = 1$ if j-th variable belongs to g-th cluster
  - Λ_c representative loadings for the g-th variable cluster

- Number of loadings: $(p \times K) \longrightarrow (G \times K)$

Some additional thoughts

- In matrix form

$$\begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,p} \end{pmatrix} = \begin{pmatrix} z_{1,1} & \cdots & z_{1,G} \\ \vdots & \ddots & \vdots \\ z_{p,1} & \dots & z_{p,G} \end{pmatrix} \begin{pmatrix} \lambda^c_{1,1} & \cdots & \lambda^c_{1,K} \\ \vdots & \ddots & \vdots \\ \lambda^c_{G,1} & \dots & \lambda^c_{G,K} \end{pmatrix} \begin{pmatrix} u_{i,1} \\ \vdots \\ u_{i,K} \end{pmatrix} + \begin{pmatrix} \varepsilon_{i,1} \\ \vdots \\ \varepsilon_{i,p} \end{pmatrix}$$

- Uniformativeness can be detected by a priori forcing a row of Λ_c to be equal to a zero vector

- Adapting distributional assumptions we obtain similar results $\rightarrow$ in particular $(x_i|Z) \sim \mathcal{N}_p(0, \tilde{\Sigma} = \tilde{\Lambda}\tilde{\Lambda}^T + \Psi)$

Model estimation - Prior specification

- A Bayesian approach to factor analysis estimation is adopted
- Standard prior distribution are assumed for Λ_c, u_i, Ψ
- Prior on the allocation matrix $Z \longrightarrow$ Product Partition Model

$$Z \sim \mathrm{PPM}(\alpha_z)$$

Correspondence between Z and the disjoint subsets clustering representation $\mathbf{c} = \{C_1, \ldots, C_G\}$ is exploited

$$\pi(\mathbf{c}) \propto \prod_{g=1}^{G} \rho(C_g) = \alpha_z^G \prod_{g=1}^{G} (|C_g| - 1)!$$

Model estimation - Algorithm

- Conjugate nature of various priors adopted allows a Gibbs updating scheme
- Metropolis-Hastings step to sample Z considering an adaptation of the allocation sampler (Nobile & Fearnside, 2007)
  - Idea: a single move attempts to reallocate a block of variables from one group to another
  - Bigger moves, faster exploration of the partition space
  - Modification aim to propose moves involving *close* clusters
- Rotational invariance and identifiability issues?

Model selection

- Number of factors K and number of variable clusters G
- Possible solutions:
  - Information criteria (AICM, BICM, BIC-MCMC...)
  - Nonparametric fashion $\longrightarrow$ infinite groups/factors
- Proposal $\longrightarrow$ ad-hoc initialization strategy
  - Idea: multi-step procedure to mimick $\tilde{\Lambda}$ structure via standard FA and model-based clustering strategy
  - Output: $(K_{\text{init}}, G_{\text{init}})$ initial guess to be used as the starting for a local search
  - Rationale: avoid intensive global search since the focus is on covariance reconstruction

Some results - Synthetic data

- $B = 200$ samples drawn from $\mathcal{N}_p(0, \Lambda_{\text{true}}\Lambda_{\text{true}}^T + \Psi_{\text{true}})$ with $n = 500$, $p = 40$ with $K_{\text{true}} = 3$ and $G_{\text{true}} = 5$

- Aim of the simulations:
  - Assess the quality of the initialization strategy
  - Evaluate covariance reconstruction and clustering of the variables

- The true model is the most selected with a tendency to overestimate K and G with the latter harder to detect

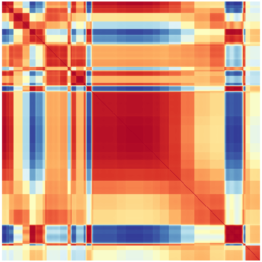
- Overestimation harmless if not beneficial in terms of `MSE` and `ARI` with G having a stronger impact on the results

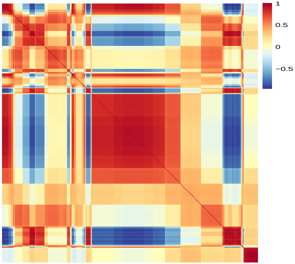
Dairy diet MIR spectroscopy data

- Milk samples from 120 cows collected weekly during summer months (from 2015 to 2017)
- Two different feeding regimens considered:
  - Pasture, cows maintained outdoors on perennial ryegrass and white clover
  - Total mixed ration (TMR), cows mantained indoors, nutrients combined in a single mix
- $n = 4320$ milk samples, $n_\mathsf{P} = 2391$ $n_\mathsf{T} = 1389$
  $p = 1060$ wavelengths in the mid-infrared region
  $\longrightarrow$ three water-absorption related spectral regions have been removed thus $p^* = 533$

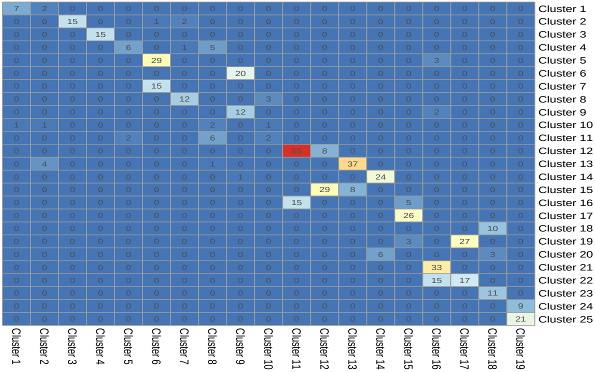
Some results - MIR data

- Initialization strategy suggests $(K_P = 4, G_P = 25)$, $(K_T = 3, G_T = 19)$
- Good reconstruction of the sample correlations
- Blocky structure as a byproduct of the clustering mechanism
  $\longrightarrow$ useful to highlight differences between spectral regions





Some results - MIR data



- Quite strong agreement between the two partitions (ARI = 0.65) $\longrightarrow$ signal about real and non-spurious clustering structures
- Indications about pronounced redundancy, grouping may be used to build new discriminating variables

Some results - MIR data

Pasture

	Fat	Protein	Lactose
Cluster−1	0.95	0.42	0.86
Cluster−2	0.97	0.69	0.97
Cluster−3	0.98	0.51	0.96
Cluster−4	0.96	0.83	0.96
Cluster−5	0.98	0.96	0.95
Cluster−6	0.98	0.94	0.96
Cluster−7	0.98	0.93	0.93
Cluster−8	0.98	0.95	0.60
Cluster−9	0.98	0.92	0.89
Cluster−10	0.93	0.89	0.01
Cluster−11	0.98	0.93	0.66
Cluster−12	0.98	0.70	0.50
Cluster−13	0.97	0.47	0.88
Cluster−14	0.98	0.87	0.96
Cluster−15	0.95	0.32	0.19
Cluster−16	0.97	0.50	0.53
Cluster−17	0.93	0.47	0.58
Cluster−18	0.98	0.73	0.90
Cluster−19	0.95	0.56	0.65
Cluster−20	0.98	0.75	0.91
Cluster−21	0.98	0.88	0.88
Cluster−22	0.96	0.84	0.88
Cluster−23	0.98	0.75	0.89
Cluster−24	0.91	0.26	0.05
Cluster−25	0.91	0.30	0.08

TMR

	Fat	Protein	Lactose
Cluster−1	0.96	0.79	0.22
Cluster−2	0.95	0.79	0.94
Cluster−3	0.96	0.76	0.98
Cluster−4	0.97	0.69	0.97
Cluster−5	0.94	0.94	0.98
Cluster−6	0.98	0.98	0.99
Cluster−7	0.97	0.97	0.98
Cluster−8	0.97	0.96	0.96
Cluster−9	0.98	0.98	0.98
Cluster−10	0.96	0.95	0.35
Cluster−11	0.98	0.83	0.81
Cluster−12	0.94	0.54	0.44
Cluster−13	0.95	0.51	0.36
Cluster−14	0.98	0.93	0.98
Cluster−15	0.96	0.68	0.81
Cluster−16	0.98	0.97	0.97
Cluster−17	0.95	0.86	0.93
Cluster−18	0.98	0.92	0.97
Cluster−19	0.89	0.53	0.29

Scale: 1, 0.8, 0.6, 0.4, 0.2, 0

- Cluster-specific predictive analysis to predict milk traits contents
- Generally good predictions, useful to understand chemical processes underneath spectral regions
- Results has to be paired with subject-matter knowledge *(galactose example)*

Concluding remarks

- The proposed method provides parsimonious summaries of high dimensional data with highly correlated variables
  $\longrightarrow$ not only spectroscopic data

- Richer insights with respect to standard FA thanks to the variable clustering mechanism

- Directions and open questions
  - It might serve as a building block for classification tools, easy to embed in a MFA framework
  - Different choices for the prior on Z?
    $\longrightarrow$ shrinkage priors
    $\longrightarrow$ exploit some *spatial* information when specifying the cohesion function

Some references

Check the paper out on arXiv
https://arxiv.org/pdf/2101.12499.pdf

Other references

- Barry, D. & Hartigan, J.A. (1992) *Product partition models for change point problems*, The Annals of Statistics, 20(1), 260–279.
- Bartholomew, D.J., Knott, M. & Moustaki, I. (2011) *Latent variable models and factor analysis: a unified approach*, John Wiley & Sons.
- Frühwirth-Schnatter, S. & Lopes, H.F. (2010) *Parsimonious Bayesian factor analysis when the number of factor is unknown*, Technical report.
- Nobile, A. & Fearnside, A.T. (2007) *Bayesian finite mixtures with an unknown number of components: the allocation sampler*, Statistics and Computing, 17(2), 147–162.